Bicluster Coherency Measures for Gene Expression Data

Mahmoud Mounir[(1)], Mohamed Hamdy[(1)], and Mohamed Essam Khalifa[(2)]

[(1)] Information Systems Department, Faculty of Computer and Information Sciences,
[(2)] Basic Sciences Department, Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt
mahmoud.mounir@cis.asu.edu.eg, m.hamdy@cis.asu.edu.eg, and esskhalifa@cis.asu.edu.eg

Abstract

Many studies have been proposed to analyze gene expression microarray data, emphasizing on the identification of genes that show related functions over only subsets of different conditions. Detection of these homogenous genes is a crucial step in this analysis. One of the main approaches to achieve this task is biclustering, which is a time-consuming process that starts with the identifying sets of genes as seeds, expanding theses seeds using heuristic searches along with a measure of coherency to assess the quality of the resulting biclusters. The identification of the suitable coherency measure is a critical task, not only affecting the expansion of initial seed biclusters, but also the final shape of them. In this paper, a number of bicluster coherency measures for gene expression data are reviewed and analyzed from both analytical and mathematical aspects to help researchers in the choice of the right measure.

Keywords: *Clustering, Biclustering, Microarrays, Gene Expression Profiles, Coherency Measures, Correlated Patterns.*

1. Introduction

The rapid evolution in the microarray technologies allow the monitoring of the expression levels of huge number of genes simultaneously under many different experimental conditions, these data are called gene expression data matrix [1,2]. Elements included in this data matrix represent the value of each gene under different experimental conditions (e.g. different tissue types, or different timestamps, etc.). To understand the relationships between these genes, the analysis of these data is a crucial step to discover and explain certain biological process [3]. To reveal the relations between different genes, machine learning approaches are applied to gene expression data [4]. Genes are said to be co-expressed or co-regulated, if these genes have high level of similarity under different subsets of experimental conditions and therefore these genes may share a common biological function or participate in the same cellular process [1]. This kind of similarity depends on the different desired types of patterns such as shifting or scaling patterns [5]. Traditional clustering techniques aim at finding different groups of genes that behave similarly under all experimental conditions or across the whole set of conditions, in other words, trying to find global relations between genes. On the other side, genes may be active or behave in the same way under only subsets of experimental conditions or participating in many different cellular process or cell functions. Therefore, the deep need arises to find data grouping technique that can find subsets of genes that share similar expression patterns under only subsets of experimental conditions, in addition to finding local relations between genes, this technique is called

biclustering. Biclustering tries to find submatrices of the original gene expression data matrix that satisfy some sort of similarity or coherency by performing clustering in both dimensions in the same time and as a result to this process, the local relations or interactions between genes are identified [6-11].

Hartigan [12] was the pioneer scientist who introduced the basic idea of biclustering as a co-clustering technique in 1970s, followed by the first application of this concept to gene expression data by Cheng and Church [13]. Biclustering problem was proved to be NP-Complete problem and more complex than traditional clustering techniques, therefore most biclustering techniques depend on optimization algorithms such as heuristic search [13-14]. Biclustering tries to identify sets of genes that have similar expression patterns under specific subsets of the experimental conditions. Starting with an (n X m) data matrix A where n denotes number of genes and m represents the number of experimental conditions. Every value in this matrix (aij) is the gene expression level for Gene (i) under a certain condition (j).

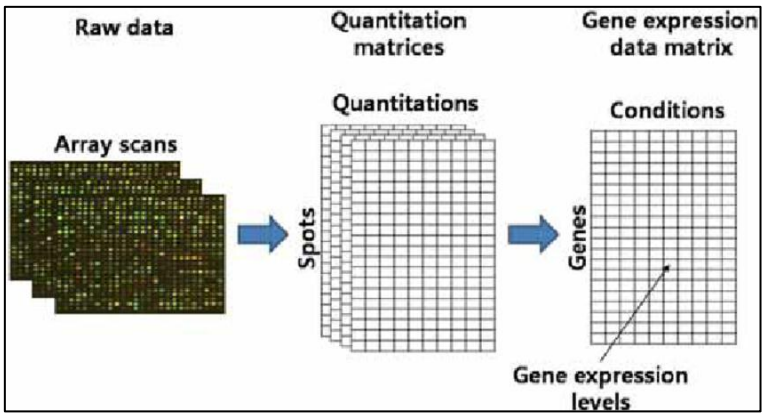

Figure 1. The Microarray gene expression data matrix

Most of the biclustering techniques starting by identifying different set of elementary or initial set of biclusters or "seeds" and continue searching and expanding these "seeds" to find optimal set of biclusters that satisfies certain coherency or quality measure. Throughout this search and expansion process, these approaches use some measure to assess the quality or coherency of biclusters and leading the search process. The development of effective and efficient coherency measure is considered an important and crucial process that has a great effect to the outcome of the resulting identified biclusters. Here we will present a taxonomy of the different coherency measures used by different biclustering approaches.

In the next section, we will present a mathematical model for the biclustering problem, following by a taxonomy of different bicluster types based on their gene expression patterns, in section 4, the different bicluster coherency measures are introduced and investigated both analytically and mathematically, finally, a conclusion and future work for this paper is introduced.

2. Mathematical modeling of biclustering

Let $\boldsymbol{X = \{G, E\}}$ represents a Gene Expression Data Such that, $\boldsymbol{G= \{G(1), G(2), \ldots, G(i)..., G(N)\}}$ and $\boldsymbol{E= \{ E(1), E(2), \ldots, E(i)..., E(M)\}}$, Where G is the set of genes and E is the set of different experimental conditions related with each gene. Now we have a Data Matrix $\boldsymbol{D}$, where

$$D = \begin{bmatrix} g(1) \\ g(2) \\ \vdots \\ g(i) \\ \vdots \\ g(N) \end{bmatrix} = \begin{bmatrix} d(1,1) & d(1,2) & \cdots & d(1,j) & \cdots & d(1,M) \\ d(2,1) & d(2,2) & \cdots & d(2,j) & \cdots & d(2,M) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d(i,1) & d(i,2) & \cdots & d(i,j) & \cdots & d(i,M) \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ d(N,1) & d(N,2) & \cdots & d(N,j) & \cdots & d(N,M) \end{bmatrix}$$

Each element in the data matrix d_{ij} represents the expression level of gene i under a specific experimental condition j.

Row number i or $\boldsymbol{g(i) = [D_{i1}, D_{i2}, ..., D_{ij}..., D_{iM}]}$ is an $\boldsymbol{(1\ X\ M)}$ vector represents the expression level of gene i under all M experimental conditions. Whereas, column number j or $\boldsymbol{E(j) = [D_{1j}, D_{2j}, ..., D_{ij}..., D_{Nj}]}$ is an $\boldsymbol{(N\ X\ 1)}$ vector represents the expression levels of all N genes under a specific experimental condition j.

Given the previous mathematical model of gene expression data matrix, we can identify a bicluster as a submatrix $\boldsymbol{B}$ of matrix $\boldsymbol{D}$ or as a subset $\boldsymbol{Y}$ of $\boldsymbol{X}$, such that

$$B = [b(i,j)] = [b_{ij}] \quad (1)$$

$$Y = \{I, J\} \quad (2)$$

Where $\boldsymbol{I}$ is a subset of $\boldsymbol{G}$ and $\boldsymbol{J}$ is a subset of $\boldsymbol{E}$.

The final goal is trying to identify several $\boldsymbol{K}$ submatrices or $\boldsymbol{K}$ biclusters such that each bicluster $\boldsymbol{B_K = [b(i,j)]}$, for each $\boldsymbol{i \epsilon I_k}$ and $\boldsymbol{j \epsilon J_k}$ that implies some measure of coherency. The Mean of the row number (i) in a certain bicluster B is defined as

$$b_{iJ} = \frac{1}{|J|} \sum_{j \epsilon J} b_{ij} \quad (3)$$

While the Mean of the column number (j) in a certain bicluster B is defined as

$$\boldsymbol{b_{iJ} = \frac{1}{|I|} \sum_{i \epsilon I} b_{ij}} \quad (4)$$

The overall mean of within a certain bicluster **B** is defined as

$$b_{IJ} = \frac{1}{|I|} \sum_{j \epsilon J} b_{iJ} = \frac{1}{|J|} \sum_{i \epsilon I} b_{Ij} \quad (5)$$

The variance within any bicluster $\boldsymbol{B}$ is computed as

$$VAR\ (B) = \sum_{i \epsilon I} \sum_{j \epsilon J} (b_{ij} - b_{IJ})^2 \quad (6)$$

There may be some additional symbols will appear throughout the paper; $\boldsymbol{\mu}$ denotes the background effect, $\boldsymbol{\alpha}$ represents the row effect, and $\boldsymbol{\beta}$ as column effect.

3. Biclusters types based on gene expression patterns

Here, we will give a brief description of different shapes of biclusters based on types of biclusters they can find. There are four major classes of biclusters: 1) Biclusters with constant values, 2) Biclusters with constant values on rows or columns, 3) Biclusters with coherent values, 4) Biclusters with coherent evolutions.

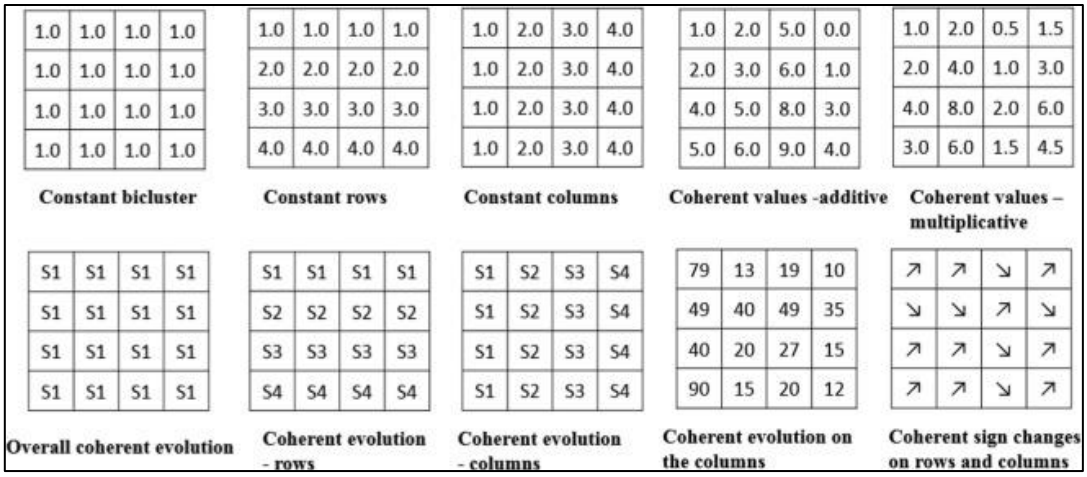

Figure 2. Different types of biclusters based on gene expression patterns

3.1 Biclustering with constant values

This type of biclusters searches for submatrices of the original gene expression data matrix that have genes with the same values under subsets of conditions.

$$B = [b_{ij}] = [constant\ value\ (\mu)] \quad (7)$$

For noise-free datasets, within each bicluster, the value of variance will be equal to zero, which can be changed in the presence of noise.

3.2 Biclusters with constant values on rows

The aim of many biclustering algorithms is the identification of biclusters that has the same expression patterns across all rows or conditions. In case of constant rows, this is a reflection of how a set of conditions are similar over a set of genes, while in case of constant columns, it reflects how a set of genes are similar across a group of conditions. The perfect constant bicluster on rows is the one where all elements in this bicluster have a constant value over its rows, where

Additive model:

$$B = [b_{ij}] = [\mu + \alpha_i] \quad (8)$$

Multiplicative model:

$$B = [b_{ij}] = [\mu\ X\ \alpha_i] \quad (9)$$

The variance of each row is calculated by

$$max\ (b(i, :)) - min(b(i,:)) \quad (10)$$

The value of variance was proven to be equal to zero in case of noise-free data, whereas, must be less than the predefined threshold (δ_i) in case of noisy data.

3.2 Biclusters with constant values on columns

Similar to the previous type, the perfect bicluster with constant values over all columns is characterized by that all elements within the bicluster have the same value over all columns, or constant expression levels across different groups of genes.

Additive model:

$$B = [b_{ij}] = [\mu + \beta_i] \quad (11)$$

Multiplicative model:

$$B = [b_{ij}] = [\mu \, X \, \beta_i] \quad (12)$$

The variance of each row is calculated by

$$max\,(b(:, j)) - min(b(:, j)) \quad (13)$$

3.3 Biclusters with coherent values and evolutions

Biclusters are identified by subsets of genes that are up-regulated or down-regulated coherently across subsets of conditions. The change in expression value over all rows or columns has the same magnitude and happens in the same direction. . This type has two models

Additive model:

$$B = [b_{ij}] = [\mu + \alpha_i + \beta_j] \quad (14)$$

Multiplicative model:

$$B = [b_{ij}] = [\mu \, X \, \alpha_i \, X \, \beta_j] \quad (15)$$

In the coherent evolutions biclusters, biclusters may represent the set of experimental conditions for different stages in a disease or cellular process that may vary in the same way or may vary by the same magnitude but in different directions. The perfect coherent evolutions bicluster is identified by that all rows have a linear order across a subset of columns or vice versa. Biclusters members may be obtained by the following formula:

$$B = [b_{ij}] = [\mu + \alpha_i \, X \, \beta_j] \quad (16)$$

3.4 Shifting and scaling expression patterns

In a shifting pattern, each column is shifted by an additive factor. A shifting pattern follows equation (17) :

$$e_{ij} = \pi_i + \beta_j \quad (17)$$

Where e_{ij} represents the expression level of gene i under specific condition j, e_{ij}, is a shifted expression of a base expression π in row i shifted by a shifting factor β in column j.

In a scaling pattern, each column is scaled by multiplicative factors. A scaling pattern follows equation (18) :

$$e_{ij} = \pi_i + \alpha_j \quad (18)$$

Here, e_{ij} is the expression level gene i under experimental condition j. It is a scaled expression level of a base expression π in row i a scaling factor α in column j.

The shifting-scaling pattern is merges shifting pattern and a scaling pattern. Each expression is shifted by a shifting factor and scaled by a scaling factor. The shifting-scaling pattern follows equation (19) :

$$e_{ij} = \pi_i \, X \, \alpha_j + \beta_j \quad (19)$$

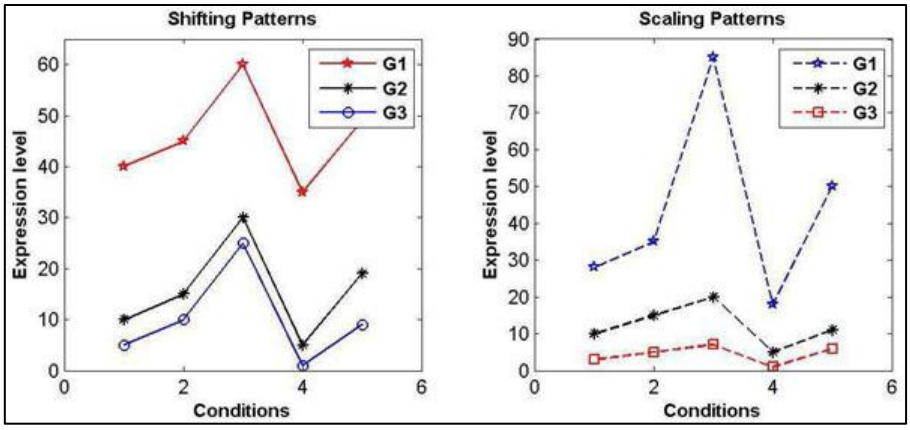

Figure 3. Shifting and scaling patterns

4. Bicluster coherency measures

In this section, we will introduce some of coherency measures that mostly or known to be used with gene expression biclusters. These measures are used to find any shape of biclusters including shifting and scaling patterns in a bicluster. But most of these measures cannot identify the perfect shifting and scaling patterns.

4.1 Variance (VAR)

The perfect biclusters identified by this measure of coherency are those that minimize the overall variance of all resulting biclusters. The overall bicluster variance cane be calculated from equation (6). Hartigan [12] used this measure as a measure of coherence in his proposed approach, where only constant value biclusters can be detected. Therefore, to detect different types of biclusters, other homogeneity measures must be used along with the variance.

4.2 Mean squared residue (MSR) and scaling mean squared residue (SMSR)

In this type of coherency measure, the good bicluster is the one that lower the mean squared residual distance among all biclusters. This is a reflection of how stronger the coherency of these biclusters is. A predefined threshold δ may be used to limit the value of MSR and in this case it is called δ-bicluster. The biclustering algorithm proposed by Cheng and Church [13] used mean squared residue combined with greedy heuristic search to find the optimal biclusters. Equation (20) is used to calculate the MSR for a bicluster B with a number of (I) rows and (J) columns.

$$MSR\ (B) = \frac{1}{|I||J|}\sum_{i\epsilon I}\ \sum_{j\epsilon J}\ (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \qquad (20)$$

A perfect bicluster is identified to be the one with MSR equals to zero. However, MSR can find only biclusters with shifting patterns, but it has some limitations that affect its ability to identify scaling patterns in gene expression data [14].

To overcome this limitation of MSR, Mukhopadhyay et al. [15] proposed an enhanced mean squared residue called scaling mean squared residue (SMSR) in equation (21) to find biclusters with scaling patterns, but it cannot detect shifting patterns identified by traditional MSR.

$$SMSR\ (B) = \frac{1}{|I||J|}\sum_{i\epsilon I}\ \sum_{j\epsilon J}\ \frac{(\ biJ\ X\ bIj - bij\ X\ bIJ)2}{(biJ\ X\ bIJ\ )2} \quad (21)$$

4.3 Relevance index coherency measure (RI)

This kind of coherency is calculated as the total relevance indices of all columns, where the relevance index of a single column is defined by equation (22)

$$RI_{Ij} = 1 - \frac{(\sigma Ij)2}{(\sigma j\ )2} \quad (22)$$

Where $\boldsymbol{\sigma_{Ij}}$ is the variance of all elements in this column or (local variance), on the other hand, $\boldsymbol{\sigma_j}$ is the variance of the whole dataset or (global variance). We can induce from the previous formula that RI gives higher value when the local variance is smaller than the global variance and therefore, for a given column, the RI is maximum if it has a local variance of zero, while the global one is not. Yip et al. [16] firstly introduced this measure of coherency in their work, however, it can only identify biclusters with constant values over rows or columns.

4.4 Correlation-based coherency measures

Gene expression microarray analysis uses this type of measure in many processes and data mining techniques especially in the clustering task. This measure emphasizes on the overall similarity of genes taking into considerations negative values of correlation as well. Here, we will give a brief description of the correlation-based coherency measures mostly use in gene expression data analysis.

4.4.1 Pearson's correlation coefficient (PCC)

Pearson correlation coefficient PCC is a measure of linear association or relationship between any two variables. It is defined as a division of the covariance of two variables by the product of their standard deviation. The value of PCC ranges from +1 to -1, where a value of +1 means a perfect direct relation or increasing linear relationship. On the other hand, -1 value implies a perfect decreasing or inverse linear relationship. All values between +1 and -1 are indicators of the degree of linear dependencies between variables except for the 0 value of PCC which implies no linear relationship between the variables. This type of coherency used efficiently on the identification of co-regulated genes over different experimental conditions [17] which helped in the identification of shifting and scaling patterns in gene expression data matrices. However, PCC has a leakage in capturing constant values biclusters or constant values over rows or columns because the zero value of standard deviation appears in the denominator. The value of PCC can be calculated between any two genes by equation (23)

$$PCC(g_1, g_2) = \frac{\sum_{j=1}^{|J|}(bi1j-bi1J)(bi2j-bi2J)}{\sqrt{\sum_{j=1}^{|J|}(bi1j-bi1J)^2}\sqrt{\sum_{j=1}^{|J|}(bi2j-bi2J)^2}} \quad (23)$$

Where b_{i1j} and b_{i2j} are the elements in rows i_1 and i_2 and column j and b_{i1J} and b_{i2J} represent the mean values of rows i_1 and i_2 respectively.

To calculate the overall PCC within a certain bicluster, PCC values for all pairs of genes in the same bicluster are calculated. An adapted version of PCC called average correlation AC were used as a measure of coherency in the work introduced by [17, 18].

$$AC\ (B) = \frac{\sum_{i1=1}^{|I|-1}\sum_{i2=i1+1}^{|I|}PCC(i1,i2)}{\binom{|I|}{2}} \quad (24)$$

4.4.2 Sub-matrix correlation score (SCS)

This type of correlation measure is defined based on Pearson coefficient correlation score by Yang et al. [19], claiming that a perfect correlation is satisfied by perfect linear relationship over rows and columns vectors separately. The score of rows and columns correlation are calculated by equations (25) and (26) respectively :

$$S_{row} = min_{i1 \in I} (S_{i1J}),\ S_{i1J} = 1 - \frac{\sum_{i2 \neq i1} |corr(xi1J, xi2J)|}{|I|-1} \quad (25)$$

$$S_{col} = min_{j1 \in J} (S_{Ij1}),\ S_{Ij1} = 1 - \frac{\sum_{j2 \neq j1} |corr(xIj1, xIj2)|}{|I|-1} \quad (26)$$

Where $corr(X_{i1J}, X_{i2J})$ and $corr(X_{I1j}, X_{I2j})$ are the Pearson correlation coefficients of all pairs of genes and conditions within the same bicluster. S_{row} and S_{col} represent the degree of correlation on rows or columns of any bicluster, respectively. The sub-matrix correlation score S(B) can be calculated as the minimum of these two score values. The perfect bicluster has a score value of zero which means that the rows or columns in this bicluster have a perfect linear relation.

$$S(B) = min\ (S_{row}(I, J)\ ,\ S_{col}(I\ ,\ J)) \quad (27)$$

4.4.3 Average Spearman's rho (ASR)

ASR was proposed by Ayadi et al. [20] and it is based on the Spearman's rank correlation which measures the statistical relationship between tow non-linear or monotonic variables. The value of ASR ranges from +1 to -1, depending on how these two variables are related. The main difference between ASR and PCC, is that ASR can detect monotonic relationships between variables while PCC can only detect only linear relationships, this implies that ASR is less sensitive to noise or outliers than PCC. The ASR can be calculated from equation (28).

$$ASR\ (B) = 2\ X\ max\ \{\ \frac{\sum_{i \in I} \sum_{j \geq i+1, j \in I} \rho ij}{|I|(|I|-1)}\ ,\ \frac{\sum_{k \in J} \sum_{I \geq k+1, I \in J} \rho KI}{|I|(|I|-1)}\ \} \quad (28)$$

Where ρ_{ij} and ρ_{IK} denote the Spearman correlation between two different genes or conditions.

4.5 Standardization-based coherency measures

The platform of gene expression microarray used to measure the expression values of genes can affect these values and can vary significantly. To make a proper comparison between genes or patterns, a standardization process of the gene expression values in each biclusters is needed to scale these values to a mutual range. Standardization is an important step that has a deep and tangible effect on the results of different biclustering techniques. Given a bicluster (B), the standardized bicluster ($B^{\backslash}$) can be obtained using the following formula:

$$b^{\backslash}_{ij} = \frac{bij - \mu gi}{\sigma gi} \quad (29)$$

Where σ_{gi} represents the standard deviation of all values of gene expression levels of gene i and μ_{gi} denotes for the mean value of row I in bicluster B. Because of using this standardization method, which reflects the relative deviation, the up/down co-regulated genes will be clear somehow. Maximal standard area (MSA) [21] is a coherency measure for biclusters that depend mainly on the previously stated gene standardization method. MSA

starting by computing the area between highest and lowest values of expression levels of a given gene in a specific bicluster under different experimental conditions. After that, the same calculation is performed to find the lowest and highest values of all genes in the bicluster. These values define a region between all conditions within the same bicluster, the area of this region represents the value of MSA. (see table 1)

The lower and higher bands are denoted by $M_j(B)$ and $m_j(B)$ and can be calculated from equation (30):

$$M_j(B) = max_i\, b_{ij}\,;\; m_j(B) = min_i\, b_{ij} \qquad (30)$$

The area of the region that represents the value of MSA is defined in equation (31):

$$MSA\,(B) = \sum_{j=1}^{|J|-1} \left| \frac{Mj(B') - mj(B') + Mj+1\,(B') - mj+1\,(B')}{2} \right| \qquad (31)$$

MSA gives higher values of coherency with less correlated genes, for a perfect bicluster, the value of MSA will be equal to zero. (see table 1)

Table 1. Summary of different bicluster coherency measures

Coherency Measure	Reference	Range of Coherency Values	Goal of Coherency Measure	Bicluster Type	Perfect Value within a Bicluster
VAR	[12]	Any Value	Variance Minimization	Constant Values Patterns	0
MSR	[13]	Any Value	Mean Squared Residue Minimization	Shifting Patterns	0
SMSR	[15]	Any Value	Mean Squared Residue Minimization	Scaling Patterns	0
RI	[16]	0 to +1	Relevance Index Maximization	Constant Values over Rows or Columns	1
PCC	[17, 18]	-1 to +1	Pearson Correlation Coefficient Maximization	Scaling and Shifting Patterns (Linear Relations)	-1 or +1
SCS	[19]	0 to +1	Sub-Matrix Correlation Score Minimization	Scaling and Shifting Patterns (Linear Relations)	0
ASR	[20]	-1 to +1	Average Spearman's Value Maximization	Scaling and Shifting Patterns (Non-Linear and Monotonic Relations)	-1 or +1
MSA	[21]	Any value	Maximal Standard Area Minimization	Scaling and Shifting Patterns	0

5. Conclusions and future work

In the recent years, biclustering techniques had a great impact in the analysis of gene expression microarray data, most of these techniques start with a random set of seed biclusters and keep searching for optimal sets of biclusters that maintain some coherency measure. In this paper we introduced some basic definition of the biclustering concept including its mathematical model, the different types of biclusters based on the gene expression pattern. After that, we presented different measures of coherency for bicluster evaluation starting by variance, then mean squared residue, relevance index, and the different types of correlation coefficient coherency measures, taking into consideration the mathematical formulas of each

measure, points of strengths and limitations facing each of them. This paper can help researchers in determining the best coherency measure to asses bicluster quality as these measures has a great effect on the bicluster search process and the last output of biclustering algorithms.

In the future, we can perform an experimental study using all previously mentioned coherency measures to evaluate the bicluster resulting from different approaches that use these measures using different performance indices in both real and synthetic benchmark datasets.

References

[1]. C. Guindalini and R. Pellegrino, "Gene Expression Studies Using Microarrays," Rodent Model as Tools in Ethical Biomedical Research, pp. 203-216, 2015.

[2]. G. Dumancas, I. Adrianto, G. Bello, and M. Dozmorov, "Current Developments in Machine Learning Techniques in Biological Data Mining," Bioinformatics and biology insights, vol. 11, 2017.

[3]. J. Chen and S. Lonardi, "Biological Data Mining," 1st edition, Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press, 2017.

[4]. J. Joung, S. Kim, S. Shin, and B. Zhang, "A probabilistic coevolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset," BMC Bioinformatics, vol. 13 , 2012**.**

[5]. J. Aguilar-Ruiz, "Shifting and scaling patterns from gene expression data," Bioinformatics, vol. 21, pp. 3840–3845, 2005.

[6]. B. Pontes , R. Giráldez, and J. S. Aguilar-Ruiz, "Configurable pattern-based evolutionary biclustering of gene expression data," Algorithms for Molecular Biology, vol. 8, no. 1, pp. 1-22, 2013.

[7]. S.C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," Computational Biology and Bioinformatics, IEEE/ACM, vol. 1, pp.24-45, 2004.

[8]. A Tanay. R. Sharan, and R. Shamir, "Biclustering Algorithms: A Survey," Handbook of Computational Molecular Biology, Edited by Srinivas AlufU, Chapman, May 2004.

[9]. P. Beatriz, G. Raúl, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," Journal of Biomedical Informatics, vol. 57, pp. 163-180, 2015.

[10]. K. Eren, M. Deveci, O. Küçüktunç, and Ü. Çatalyürek, "A comparative analysis of biclustering algorithms for gene expression data," Briefings in Bioinformatics, Oxford Journals, vol. 14, no. 3, 2012.

[11]. M. Mounir and M. Hamdy, "On biclustering of gene expression data," IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 641-648, 2015.

[12]. J. A. Hartigan, "Direct clustering of a data matrix," Journal of the American Statistical Association, vol. 67, pp. 123-129, 1972.

[13]. Y. Cheng and G. Church, "Biclustering of expression data," ISMB, vol. 8, pp. 93–103, 2000.

[14]. H. Ben Saber and M. Elloumi, "A New Study on Biclustering Tools, Biclusters Validation and Evaluation Functions," International Journal of Computer Science & Engineering Survey (IJCSES), vol. 6, no. 1, 2015.

[15]. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, " A novel coherence measure for discovering scaling biclusters from gene expression data," Journal of Bioinformatics and Computational Biology, vol. 7, pp. 853–868, 2009.

[16]. K. Yip, D. Cheung, and M. Ng, "Harp: A practical projected clustering algorithm," IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 1387–1387, 2004.

[17]. Y. Zhang, J. Xie, J. Yang, A. Fennell, C. Zhang, and Q. Mam, "QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data," Bioinformatics, vol. 33, no. 3, pp. 450-452, 2017.

[18]. S. Mitra, R. Das, H. Banka, and S. Mukhopadhyay, "Gene interaction—an evolutionary biclustering approach," Information Fusion, vol. 10, pp. 242–249, 2009.

[19]. W. Yang, D. Dai, and H. Yan, "Finding correlated biclusters from gene expression data," IEEE Transactions on Knowledge and Data Engineering, vol. 2, pp. 568–584, 2011.

[20]. W. Ayadi, M. Elloumi, and J. Hao , "A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data," BioData mining, vol. 2, 2009.

[21]. R. Giraldez, F. Divina, B. Pontes, and J. S. Aguilar-Ruiz, " Evolutionary search of biclusters by minimal intrafluctuation," IEEE International Fuzzy Systems Conference, pp. 1-6, 2007.